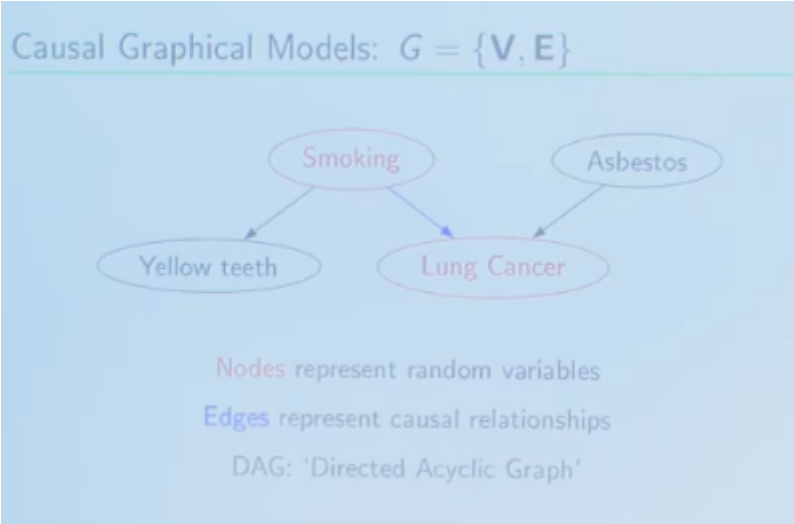
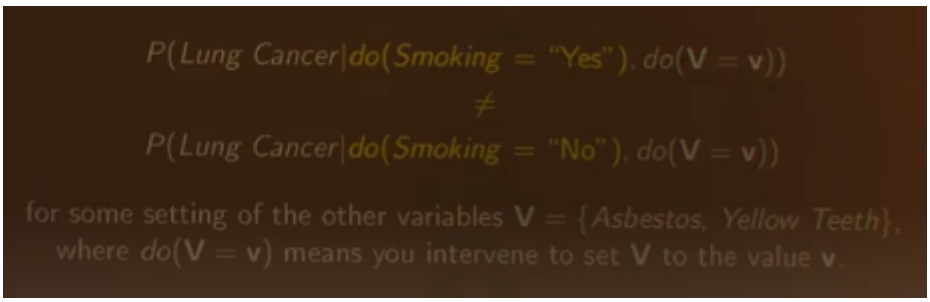
Elizabeth Silver-- Causality and Casual Discovery

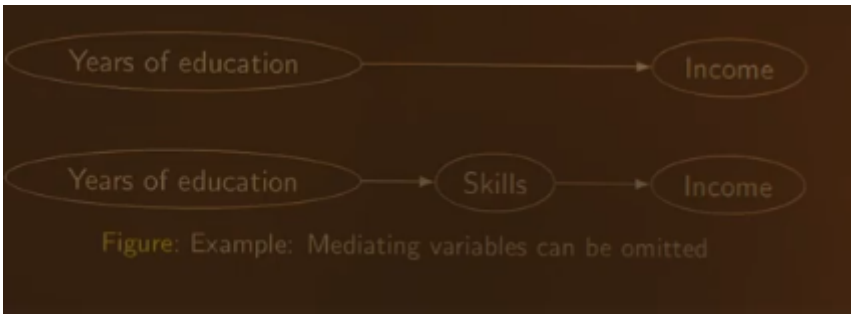
https://www.youtube.com/watch?v=51n9XsDSNps

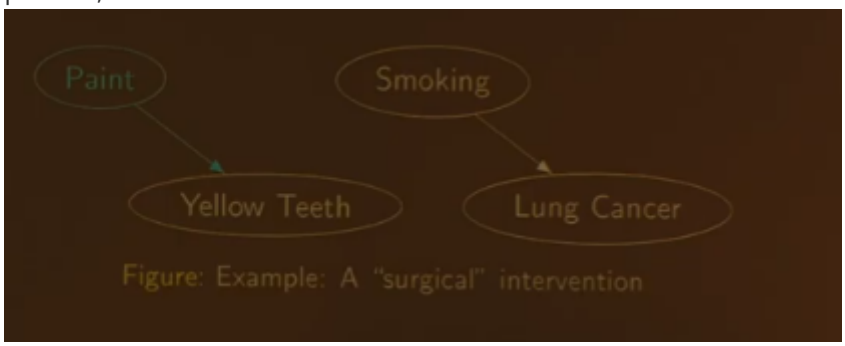
- **Input** data to causal discovery methods: RVs measured across n observations.
- **Output** of causal discovery methods: causal relationships amongst the random variables, represented by SEM:
  - $Smoking = f_1(e_1)$
  - $Asbestos = f_2(e_2)$
  - $YellowTeeth = f_3(Smoking, e_3)$
  - $LungCancer = f_4(Smoking, Asbesto, e_4)$
- If Lung Cancer is "caused" by Smoking, then Smoking will be an argument in its structural equation (f_4)
- Each equation in the SEM is a function of the causes and an error term.
- 

- Causal relationships can also be represented in graphical form (Directed Acyclic Graph DAG). Nodes are RVs, edges are causal relationships. DAGs are just short-form for SEMs.
- Notice that since Smoking has no causal parent (f_1 is only a function of the error term), it has no arrows pointing to it.
- The edges are to be learnt by causal discovery methods.
- Do operator: you are not just observing the probability of cancer from smoking, you are actively changing the probability of cancer by forcing (intervening) someone to smoke.

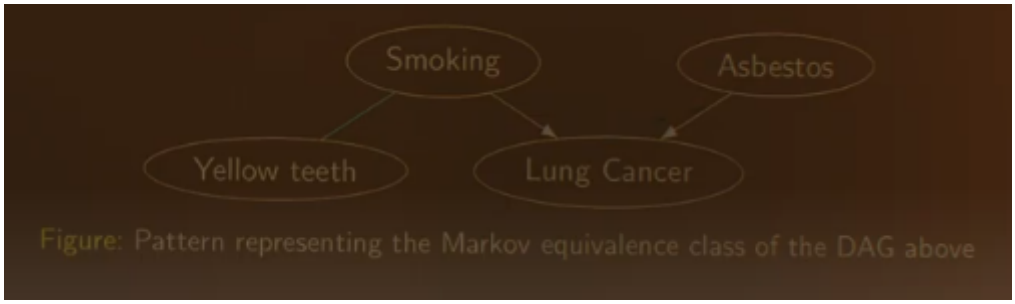


- Correlation runs in both direction: X is correlated to Y; Y is correlated to X
- Causation might only run in one direction: X causes Y; Y does not cause X.
- Smoking causes an increase in the probability of getting cancer. It is probabilistic rather than deterministic since not EVERYONE who smokes get cancer.
- Observing yellow teeth is correlated with cancer. Intervening (paint someone's teeth yellow) does not cause cancer $P(Cancer|do(yellow\_teeth)) = 0$



- Since years of education causes income through skills, I have no direct arrow from education to income in the bottom graph.
- Interventions add a node and an edge to the model.
- **Hard (surgical) interventions** break all other edges into the child of the intervention node (yellow teeth in our case below). For example, the link between smoking and yellow teeth is broken. Smoking NO LONGER have effect on yellow teeth since I am intervening on yellow teeth by painting them.
- I also need to remove the links between the intervention RV and its causal parent since we are setting the value of the intervention RV ourselves. Suppose Paint has some causal parents, I need to severe their links to Paint.
- 

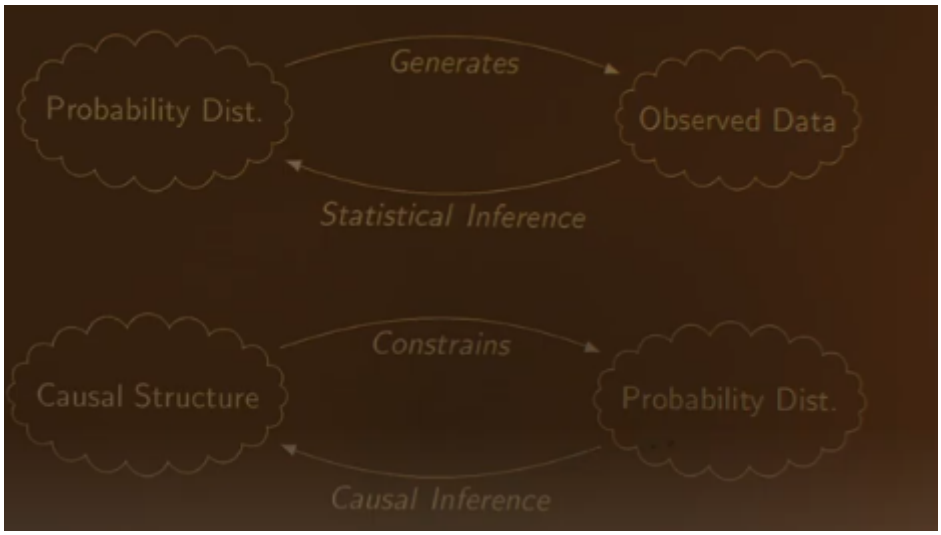
- **Soft intervention** allows the thing you are intervening (ex: yellow teeth) to be affected by things OTHER than the intervention (Paint). In this case, I keep the link between Smoking and Yellow teeth.
- Since Smoking causes both Yellow teeth and Cancer, Smoking is a confounder variable in this model. The effect of Smoking on Cancer might be absorbed by the effect of Yellow Teeth and Cancer (spurious relationship) if Cancer is omitted from the model.
- The output of causal discovery algorithms can have uncertainty:



- Notice that the algorithm identifies 3 edges, but it doesn't know the causal direction of smoking and yellow teeth.

- The goal of causal discovery approaches is: learning from data as much about the casual structure of the RVs as possible, and represent the remaining uncertainty accurately (unknown causal directions)-- be clear about what we know and don't know.

Causal Structural Learning

- **Step 1:** I want to learn from data what the edges are amongst the RVs.
    - In other words, learn the arguments in each of the structural equations in the SEM.
- Step 2: Once you know what the edges are, then you train the SEM.
- 

Constraint Type	Distributions
Conditional Independences	All
Vanishing determinants of partial covariance matrices	Linear Gaussian with unobserved confounders
Unequal dependence on residuals	Non-linear additive noise; or Linear non-Gaussian
⋮	⋮

- Main constraint type is conditional independencies.

Conditional independence constraints: the (Local) Markov Condition

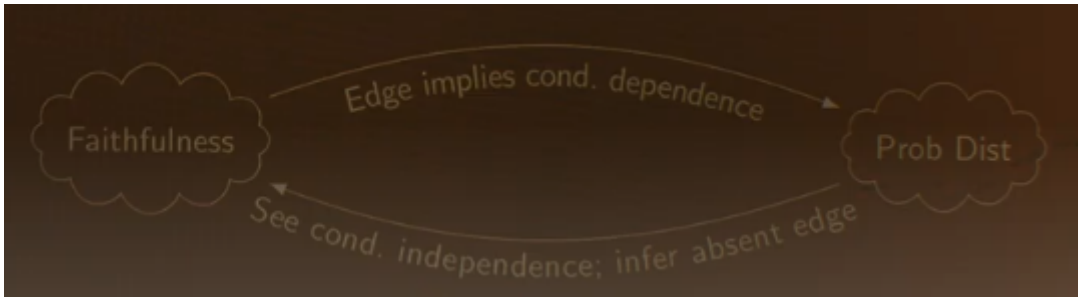
- Basically says that for any variable X, if you condition on its causal parent (any variable with arrows that points to X), then all of its non-descendants (any RV that X DOES NOT points to) are independent on X.
- If there is no edge between two RVs (say X and Y) in the causal model, then there is some conditional independence between X and Y. Lack of edge implies cond. independence.
- If I know (from data) there is conditional dependence between two variables, then we infer there must be edge between them in the causal model.

Failure of Markov Condition

- Markov condition fails if there is causes that are omitted from the model.
- Suppose in my model, I have Yellow Teeth and Cancer, but I omitted Smoking.
- I **know** that Yellow Teeth and Cancer has no causal relationship. According to Markov condition, I expect no edge in the true causal model.
- Since I omitted Smoking, a true cause of Cancer, from the model, and that Smoking causes BOTH Cancer and Yellow Teeth (and thus is a confounder in the model), then Markov condition is violated (note: specifically, the "Causal sufficiency" condition, which requires no confounders to be omitted from the model, is violated). In this case, because we will have dependence between Yellow Teeth and Cancer (spurious relationship), and Markov condition will incorrectly infer an edge between Yellow Teeth and Cancer.
- NOTE: even if I do not observe Smoking, I can include in my model to avoid violating the Markov condition.

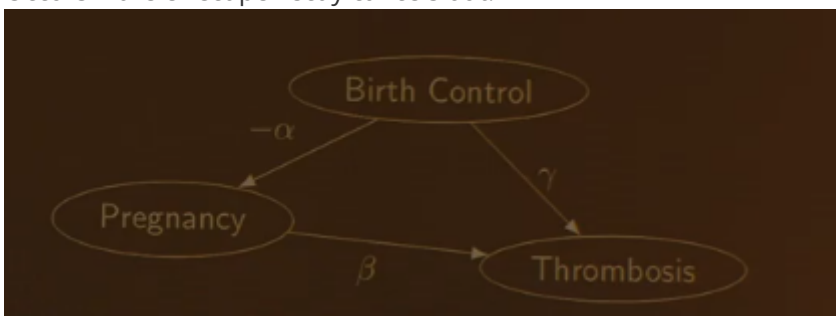
Faithfulness (the inverse of Markov Condition)

- if there is an edge between two RVs (say X and Y) in the causal model, then there IS some conditional dependence between X and Y (can be conditional on ANY other variable in the model). Edge implies cond. dependence.
- If you know (from dataset) there is cond. independence between variables, then you infer no edge in the causal model.



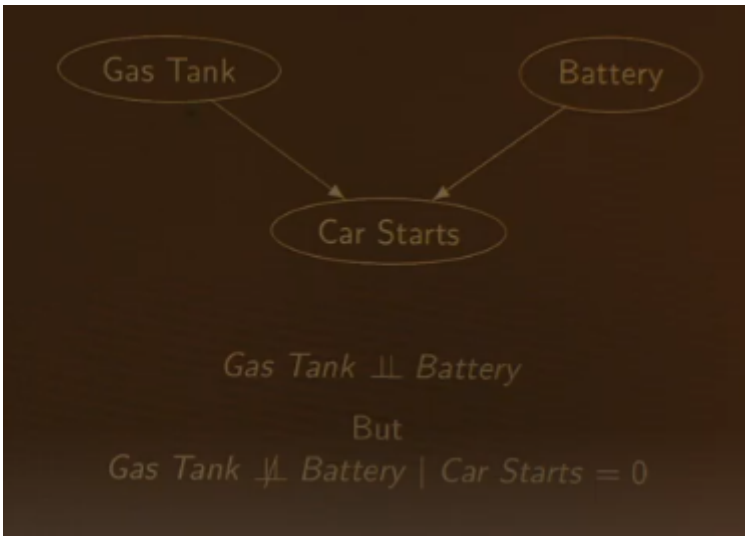
- This condition is important because it allow us to "trim down" our model and allows the algorithm to learn a sparse model.

Failure of Faithfulness (rare)

- Occurs if the effect perfectly cancels out.
- 

- Birth control affects Thrombosis (blood clots) directly and indirectly.
- I want to know γ , the direct effect.
- If α * β exactly equals γ, then faithfulness condition is violated.
- Birth control is independent from Thrombosis in this case (even tho this is not true in the actual casual model).

Example (V Structure)



- A child has two causal parents.
- The two parents are independent.
- But as soon as you condition on the child, the parents become dependent.
- Gas Tank and Battery (parents) are independent from each other.
- But given that car (child) does not start, if I know Gas tank has full tank, then I can infer Battery is dead. So Gas tank provides new information that affects the probability of battery=0. Thus, Gas Tank and Battery are dependent.
- This is called selection bias or collider bias.

Markov Equivalence Class

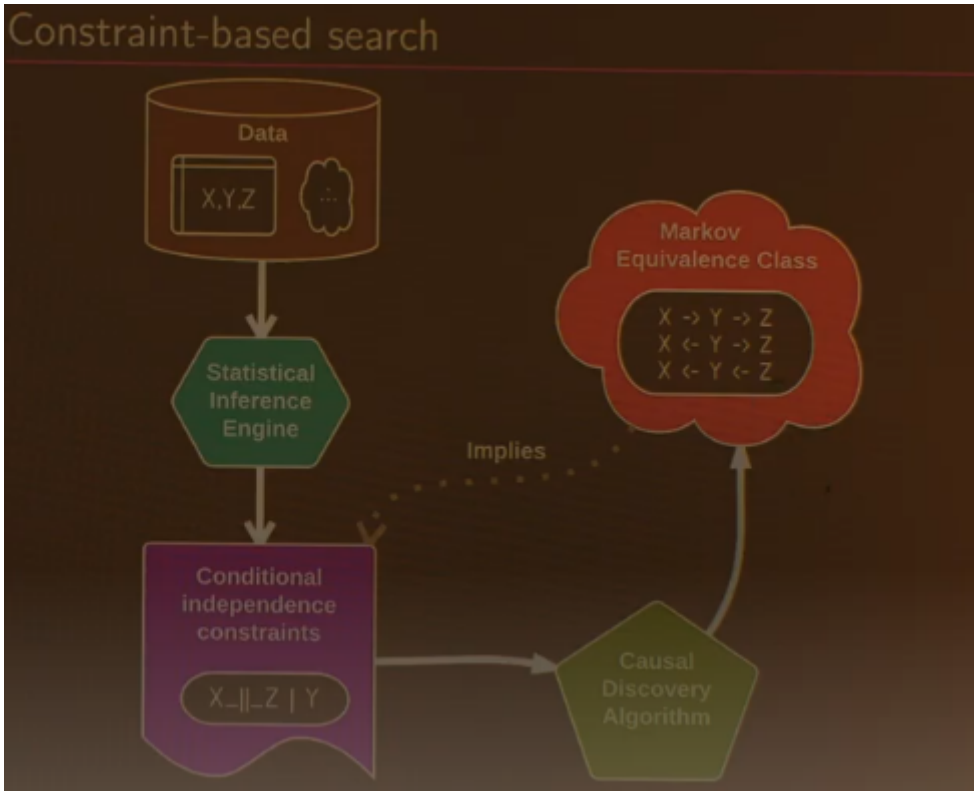
- A-->B--->C
- This structure gives us one cond. independence.
- Recall Markov Condition: for any variable X, if you condition on its causal parent (any variable with arrows that points to X), then all of its non-descendants (any RV that X DOES NOT points to) are independent on X.
- In this case, for variable C, conditional on B, A is independent since A is not C's descendent.
- $A \perp\!\!\!\perp C|B$
- The above cond. independence is implied by two other "equivalent" models:
  - A<--B<---C
  - A<---B--->C
- So if I "know" (from my sample) that one pair of variable is independent (say A and C), and all other variables are dependent. I can infer no edge between A and C, but I cannot infer the direction/orientation of the two edges since there are three possible causal model.
- The three possible/equivalent models are known as the "Markov Equivalence Class", they can be represented by:
  - A --- B --- C
  - Note that this model has some uncertainty in the causal direction.
  - It is called a CPDAG (completed partially DAG)

True DAG	$A \to B \to C$	$A \to B \leftarrow C$
Observed CIs	$A \perp\!\!\!\perp C \mid B$	$A \perp\!\!\!\perp C \mid \emptyset$
Set of DAGs in MEC	$A \to B \to C$ $A \leftarrow B \leftarrow C$ $A \leftarrow B \to C$	$A \to B \leftarrow C$
CPDAG	$A - B - C$	$A \to B \leftarrow C$

D-separation: A recipe linking CIs to graph constraints

- Check this out!
- Start with a graph, then d-separation tells us all of the cond. independencies imply by the graph.
- If two RVs in the graph is d-separated (a set of conditions met), then they are cond. independent.

Casual Discovery: Constraint based search

- 

- Input: RVs measured across n observations.
- Statistical inference: find all possible conditional independencies implied by the data.
- Using the constraints, the causal discovery algorithm find the Markov Equivalence class that is consistent with/implies the cond. independence constraints.
- User can specify any conditional independences test that you like to the statistical inference engine.
- Cons: very large search space for the Markov Equivalence class.

PC Algorithm

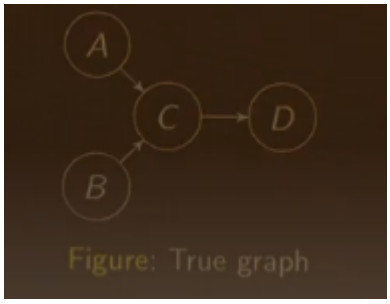
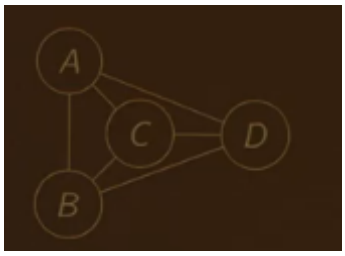
- 

Figure: True graph

- PC starts with a complete graph: everything's connected to everything else.
- 

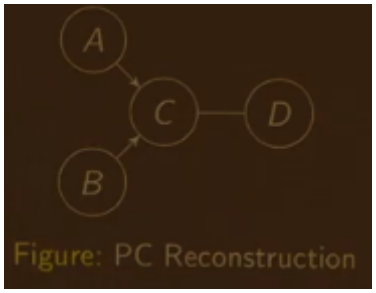
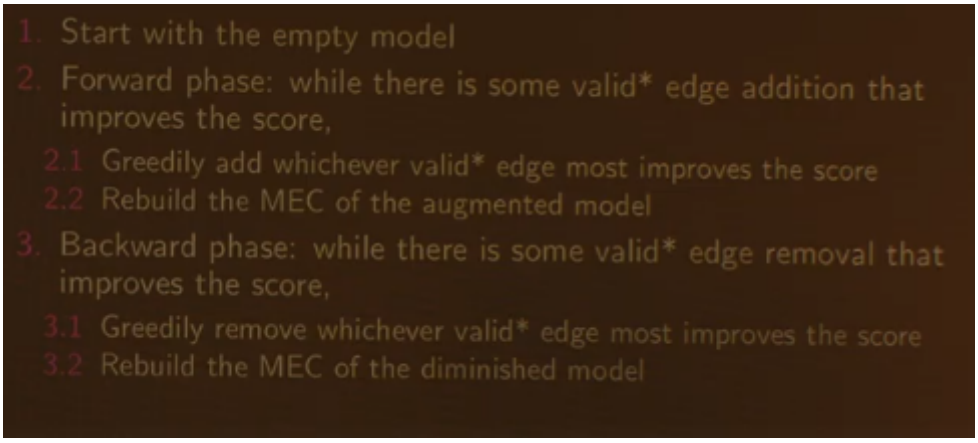
- Test "zero-order conditional independences". We find from the data that A indep B (conditional on nothing). So we remove link between A and B by Faithfulness. And by Markov, we have to keep all other edges.
- Next, find the first order conditional independencies (X indep Y | Z). I find:
  - A indep D |C
  - B indep D |C
- This is expected based on the Markov Condition (C "screens off" the connection between A-D and B-D). The conditioning set is limited since all the screening off is done by conditioning on the parents of each node. In short, I am only conditioning on a subset of variables. As the links gets sparser, the number of independence tests I need to do decreases.
- Remove link between A-D and B-D by Faithfulness. Keep everything else linked by Markov.
- No other high-order conditional independences observed.
- 

Figure: PC Reconstruction
- Look for V structures. C has two parents A and B. If A is dependent on B (conditional on C), then I can determine the causal direction of A-C and B-C. Due to the conditional dependence of A and B, I know A causes C and B causes C.
- Orient C-->D. Otherwise, I would have found A dep D |C (which I didn't).

Pro: PC is very efficient and fast on sparse model (not a lot of nodes and connection).

Con: Errors will propagate. If it messes up in the beginning, it generates other errors downstream. Also dependent on "good" independence tests.

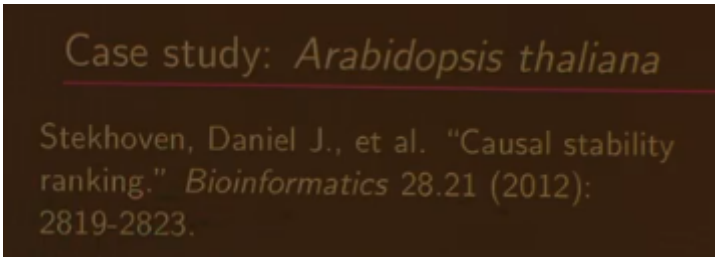
Score based Search

- Sample a model from each possible Markov Equivalence class, fit data to that model, compute a performance metric (goodness of fit score like BIC).
- Choose model with highest score.
- Suffers from the same problem as constraint-based search in that the search space is large.
- I can artificially reduce the search space by placing constraints like "no node will have more than 4 parents".
- Alternatively, use a greedy approach so that I won't have to find score for all the models. I can build the causal model like I build a decision tree. I add connections that result in largest score. (Check out Greedy Equivalence Search GES)
  - 

- MEC: Markov Equivalence Class

How do we validate our results

- Hard to find a good validation set for causal discovery. I need to test on a set of experimental outcomes for causal inference as oppose to a set of observed outcomes for ML.
- Space of potential experiments is large-- I can intervene on any variables. The intervention values can also vary for a variable.
- Alternatives:
  - Get theoretical results.
  - Validate on simulated data. "Manually" create a generative model (where I know the true causal structure). From this model I generate some synthetic data from it.
  - Take the synthetic data and see if you can learn back the generative model from it using your algorithm.
  - https://openai.com/blog/generative-models/
  - Cons: Hard to generate synthetic data where some assumptions that the casual inference algorithm depends on will be violated in a realistic way.
  - Almost every case in reality will violate some of these assumptions (ex: causal sufficiency, faithfulness). We are interested in how robust the algorithms are to the violations of these assumptions.



- Example of where validation done well.
- Used a variant of PC algorithm to find gene expression that affects flowering time.
- Can help prioritize experiments. If the causal discovery algorithm identifies the top 25 genes that causally affects flowering time, then this can save the biologist lots of time (don't have to run experiments on 10000s of genes).